npg

ORIGINAL ARTICLE

Novel organization of aromatic degradation pathway genes in a microbial community as revealed by metagenomic analysis

Hikaru Suenaga[1], Yoshinori Koyama[1], Masatoshi Miyakoshi[2], Ryo Miyazaki[2,5], Hirokazu Yano[2,6], Masahiro Sota[3], Yoshiyuki Ohtsubo[2], Masataka Tsuda[2] and Kentaro Miyazaki[1,4]

[1]*Institute for Biological Resources and Functions, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba Central 6, Tsukuba, Ibaraki, Japan;* [2]*Department of Environmental Life Sciences, Graduate School of Life Sciences, Tohoku University, Katahira, Sendai, Japan;* [3]*Integrated Technology Development Team, R & D Center, Nagase Co. Ltd., Nishi-ku, Kobe, Japan and* [4]*Department of Medical Genome Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Tsukuba Central 6, Tsukuba, Ibaraki, Japan*

Several types of environmental bacteria that can aerobically degrade various aromatic compounds have been identified. The catabolic genes in these bacteria have generally been found to form operons, which promote efficient and complete degradation. However, little is known about the degradation pathways in bacteria that are difficult to culture in the laboratory. By functionally screening a metagenomic library created from activated sludge, we had earlier identified 91 fosmid clones carrying genes for extradiol dioxygenase (EDO), a key enzyme in the degradation of aromatic compounds. In this study, we analyzed 38 of these fosmids for the presence and organization of novel genes for aromatics degradation. Only two of the metagenomic clones contained complete degradation pathways similar to those found in known aromatic compound-utilizing bacteria. The rest of the clones contained only subsets of the pathway genes, with novel gene arrangements. A circular 36.7-kb DNA form was assembled from the sequences of clones carrying genes belonging to a novel EDO subfamily. This plasmid-like DNA form, designated pSKYE1, possessed genes for DNA replication and stable maintenance as well as a small set of genes for phenol degradation; the encoded enzymes, phenol hydroxylase and EDO, are capable of the detoxification of aromatic compounds. This gene set was found in 20 of the 38 analyzed clones, suggesting that this 'detoxification apparatus' may be widespread in the environment.
The ISME Journal (2009) **3,** 1335–1348; doi:10.1038/ismej.2009.76; published online 9 July 2009
Subject Category: microbial ecology and functional diversity of natural habitats
Keywords: aromatic compounds; degradation pathways; extradiol dioxygenase; metagenomics; mobile genetic element

Introduction

The ability to use various aromatic compounds to serve as sources of carbon and energy is widespread

Correspondence: K Miyazaki or H Suenaga, Institute for Biological Resources and Functions, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba Central 6, 1-1-1 Higashi, Tsukuba, Ibaraki 305-8566, Japan.
E-mails: miyazaki-kentaro@aist.go.jp or suenaga-hikaru@aist.go.jp or M Tsuda, Department of Environmental Life Sciences, Graduate School of Life Sciences, Tohoku University, 2-1-1 Katahira, Sendai 980-8577, Japan.
E-mail: mtsuda@ige.tohoku.ac.jp
[5]Current address: Department of Fundamental Microbiology, University of Lausanne, Lausanne, Switzerland
[6]Current address: Department of Biological Sciences, University of Idaho, Moscow, ID, USA
Received 3 April 2009; revised 2 June 2009; accepted 2 June 2009; published online 9 July 2009

in bacteria, and in the natural environment these bacteria contribute greatly to the breakdown of aromatic compounds and to the global carbon cycle (Wittich, 1998; Esteve-Núñez *et al.*, 2001; Furukawa *et al.*, 2004). The majority of reported bacterial aromatic degradation processes are aerobic (Gibson and Harwood, 2002) and comprise series of enzymes that are usually categorized as either 'upper'- or 'lower'-pathway enzymes (Williams and Sayers, 1994). Generally, the enzymes in the upper pathway transform aromatic compounds to aromatic vicinal diols. This initial hydroxylation step is performed by a monoxygenase or dioxygenase that incorporates an oxygen atom(s) into the aromatic ring (Gibson and Parales, 2000). The second enzyme in the upper pathway is dihydrodiol dehydrogenase, which catalyzes the conversion of dihydrodiol to dihy-

droxy compound. In the lower pathway, the resulting dihydroxylated aromatic compounds are transformed into ring-cleavage products by either extradiol dioxygenases (EDOs) or intradiol dioxygenases. The subsequent metabolic steps are referred to as *meta*- or *ortho*- pathways. Of these dioxygenases, EDOs can easily be identified due to their yellow pigmentation of the products and have been extensively studied (Eltis and Bolin, 1996). The ring-cleavage products are further degraded into compounds that can enter the tricarboxylic acid cycle.

The genes encoding upper- and lower-pathway enzymes are often clustered into operons, and the varied combinations of upper and lower pathways provide diverse functions. Point mutations in conjunction with various gene rearrangements (for example, insertion, deletion, duplication and inversion) further promote the functional diversity of genes and gene clusters (van der Meer *et al.*, 1992). The presence of foreign compounds can lead to the selection of mutant bacteria that are capable of metabolizing them. Aromatic catabolism genes are often harbored by mobile genetic elements (for example, plasmids, transposons and integrative and conjugative elements) that effectively disseminate the catabolic traits to phylogenetically diverse bacteria (Tsuda *et al.*, 1999; Top and Springael, 2003; Nojiri *et al.*, 2004).

However, because the generalizations described above come only from studies of those bacteria that can be cultivated under standard laboratory conditions, they are probably biased. As the vast majority (>99%) of bacteria in the natural environment are difficult to culture in the laboratory, and are thought to differ greatly from known cultured ones (Amann *et al.*, 1995), the not-yet cultivated bacteria may possess novel genes, enzymes or pathways. In addition, the aromatic compound-degrading bacteria thus far analyzed were isolated on the basis of their ability to utilize aromatics as their sole carbon and energy sources. This approach can isolate only those bacteria that possess all of the genes necessary to completely degrade aromatic compounds. Efforts have been made to directly obtain genes and pathways for aromatics catabolism using methods independent of bacterial culture (Galvao *et al.*, 2005; Mohn *et al.*, 2006). For example, Junca and Pieper (2004) and Sipilä *et al.* (2008) relied on a PCR-based approach to assess the diversity of EDO genes in the environment. However, these sequence-driven approaches still suffer from bias, in this case from the primer sequences selected for PCR.

To overcome these drawbacks, we previously employed a function-driven metagenomic approach (Suenaga *et al.*, 2007). Using environmental DNA samples prepared from an activated sludge used to treat coke plant wastewater containing various aromatic compounds, we constructed a metagenomic fosmid library in *Escherichia coli*. Using catechol as a substrate, we screened the library for EDO activities.

Of the resulting 91 fosmid clones expressing EDO activities, 38 were subjected to DNA sequencing, which resulted in the identification of 43 EDO genes. Comparison of the amino acid sequences of these EDOs with those of previously reported EDOs revealed that more than half (25) of our newly identified EDOs were from novel subfamilies: I.1.C (2 clones), I.2.G (20 clones), I.3.M (2 clones) and I.3.N (1 clone). This result suggested a striking difference between the aromatics-degradation genes in our target microbial community and those that were previously identified in known bacteria through culture-dependent approaches. In this study, we investigated the organization of the degradation pathway genes in the 38 sequenced fosmid clones.

Materials and methods

DNA sequencing and gene annotation
Shotgun DNA sequencing was performed at the Dragon Genomics Center (Mie, Japan) as described previously (Suenaga *et al.*, 2007). Sequences were assembled using Paracel Transcript Assembler software (Paracel Inc., Pasadena, CA, USA). Open reading frames (ORFs) were identified using Artemis software (Rutherford *et al.*, 2000), which was also used to calculate $\%(G + C)$ content. BLAST searches (Altschul *et al.*, 1997) were carried out against the DDBJ/EMBL/GenBank DNA databases, and the final ORF annotations were made manually. Conserved amino acid sequences were identified using the InterProScan (Quevillon *et al.*, 2005) and Pfam (Finn *et al.*, 2008) programs. A tRNA gene search was carried out using the tRNAscan-SE program (Lowe and Eddy, 1997). Self-organization map (SOM) analysis (Abe *et al.*, 2003) was performed using each 5-kb DNA window of inserts as a query. Pairwise nucleotide sequence comparisons were made and visualized using the GenomeMatcher program (Ohtsubo *et al.*, 2008). The nucleotide sequences reported here have been deposited in GenBank/EMBL/DDBJ under accession numbers AB266111–AB266148 and AB471759–AB471780.

Functional characterization of selected degradation gene clusters
PCR-mediated subcloning was used to clone some of the degradation pathway genes from fosmids 2C5 and 9C3 and from the fosmids carrying I.2.G subfamily EDOs. PCR was carried out with oligonucleotide primers (Supplementary Table S1). The amplicons were digested with *Xba*I and cloned into the *Xba*I site (2C5) or *Xba*I–*Hin*cII site (9C3 and I.2.G EDO genes) of pUC18 so that transcription of the inserted genes was under the control of the vector-derived *lac* promoter. *E. coli* JM109 cells carrying the recombinant plasmids were grown at 37 °C in Luria–Bertani (LB) medium (Sambrook and Russell, 2001) containing $50 \mu g \, ml^{-1}$ ampicillin and $0.1 \, mM$ isopropyl-β-D-thiogalactoside.

Results and discussion

General features of the metagenomic fosmid inserts
Shotgun DNA sequencing was employed for 38 fosmids. Single contigs were obtained for 22 fosmids. Table 1 summarizes the general features of the fosmid inserts, which together amounted to 1.5 Mb. The insert fragments ranged from 29.1 kb to 41.6 kb in length, from 53.9% to 69.3% in G+C content, and from 26 to 45 in the number of ORFs. In total, there were 1317 ORFs.

As no rRNA genes were found, the microbial origins of the insert fragments were predicted by SOM (self-organization map) analysis, which suggested that the majority of the metagenomic fragments originated from phylum *Proteobacteria*. This finding might represent an initial screening bias resulting from the use of *E. coli* as the host. The predicted functions of the 1317 ORFs are summarized in Table 2, and a complete list of ORFs describing their position, %(G+C) content, predicted function, predicted microbial origin and closest homologs in public databases is shown in Supplementary Table S2. When the 1317 ORFs were classified on the basis of their predicted function and compared with those identified in purely cultivated aromatics-utilizing bacteria, the metagenomic DNA fragments displayed a lower proportion of aromatics-degrading genes and a higher proportion of genes for replication/maintenance/processing of DNA, conjugative transfer and DNA transposition and site-specific recombination than those in the genomes of cultivated strains. The details are described in the Supplementary information and in Supplementary Tables S3 and S4.

Diverse genetic organization of aromatic degradation genes in metagenomic DNA
The genetic organization of representative fosmid inserts is depicted in Figure 1, and the details of the ORFs for all fosmid inserts are shown in Supplementary Figures S1, S2 and S3. Most EDO genes were flanked by genes or gene clusters predicted to be involved in the degradation of aromatic compounds. These gene products showed moderate to significant levels of identity with well-studied catabolic enzymes from isolated bacteria; these enzymes included Dmp proteins from the phenol-catabolizing plasmid pVI150 of *Pseudomonas* sp. CF600 (Shingler *et al.*, 1992), Phe proteins from the phenol-utilizing strain *Bacillus thermoglucosidasius* A7 (Duffner *et al.*, 2000), Tdn proteins from the aniline-catabolizing plasmid pTDN1 of *P. putida* UCC22 (Fukumori and Saint, 2001), Nah proteins from the naphthalene-catabolizing plasmid NAH7 of *P. putida* G7 (Sota *et al.*, 2006a) and Nag proteins from the naphthalene-utilizing strain *Ralstonia* sp. U2 (Zhou *et al.*, 2001). Schematic drawings of these pathways are shown in Figures 2–5. Notably, (i) these known aromatics degraders generally carry one or two well-organized gene clusters that allow the complete degradation of aromatic compounds to pyruvate and acetyl-CoA, and (ii) the gene clusters are usually arranged according to the order of the reactions. We refer to these representative pathways identified in isolates in discussing the properties of our fosmid fragments below.

Fosmids carrying I.2.A EDO genes. A representative EDO belonging to the I.2.A subfamily is catechol 2,3-dioxygenase. I.2.A EDOs are frequently found in *Pseudomonas* spp. and are involved in the degradation of single-ring aromatic compounds (for example, phenol and toluene) and naphthalene (Eltis and Bolin, 1996). The first step in bacterial phenol degradation is, in most cases, monohydroxylation of phenol at the *ortho*-position to form catechol. This reaction is performed by a phenol hydroxylase comprising a single component (such as PheA from *B. stearothermophilus* BR219 (Kim and Oriel, 1995)), two components (such as PheAB from *B. thermoglucosidasius* A7 (Duffner *et al.*, 2000)), or multiple components (such as DmpKLMNOP from pVI150 (Shingler *et al.*, 1992)) (Figures 2 and 5).

Plasmid pVI150 carries a catabolic *dmp* operon containing 15 structural genes that encode a multi-component-type phenol hydroxylase (DmpKLMNOP) and catechol 2,3-dioxygenase (DmpB). The complete degradation of the 2-hydroxymuconic semialdehyde formed by the DmpB reaction is carried out by the enzymes encoded by *dmpCDEHFGI*. Among our metagenomic fosmid inserts, I.2.A EDO genes were found in five fosmids that were classified into three subtypes (2C5, 3A2 and 5F10) depending on their genetic organization (Figure 1 and Supplementary Figure S1).

The fosmid 2C5 appears to contain *dmpKLMNOPQBCDEHFGI*, a complete gene cluster comprising ORF2C5-7 to −13 and ORF2C5-15 to −22; this cluster encodes polypeptides calculated to be 20–78% identical to the Dmp proteins of pVI150 (Figures 1 and 2). Upstream of this gene cluster is ORF2C5-23, which encodes a deduced polypeptide that is 61% identical to pVI150-encoded DmpR, a positive transcriptional regulator of the *dmp* operon (Figure 2). When the ORF2C5-15 to −22 cluster, presumably encoding phenol hydroxylase and EDO, was subcloned into pUC18, *E. coli* JM109 cells carrying the recombinant plasmid produced 2-hydroxymuconic semialdehyde (detected by yellow coloration) on LB agar plates containing phenol, confirming the activities of the two enzymes.

The predicted products of ORF3A2-1, −3, −5 and −15 to −20 from fosmid 3A2 are 35–79% identical to two components of phenol hydroxylase (DmpOP) (ORF3A2-1 and −3), catechol 2,3-dioxygenase (DmpB) (ORF3A2-5) and the DmpCDEFGH proteins (ORF3A2-15 to −20) of pVI150 (Figure 1). The deduced gene product of ORF3A2-14, which was located upstream of *dmpCDEFGH*, is 43% identical to the Orf0 protein, a GntR-type transcriptional

Table 1 Summary of thirty-eight environmental DNA fragments

EDO subfamily	Fosmid name	Number of contigs	Size (kb)	G+C content (%)	Number of ORFs	Phylogenic affiliation by SOM analysis	Putative degradation genes[a]	Remarks
I.2.A	2C5	2	35.1	53.9	31	α/β/γ-Proteobacteria	*dmp*RKLMNOPQ**B**-C*DEHFGI*, *phe***BC**	complete *dmp* gene cluster and two EDO genes (I.2.A and I.2.G subfamilies)
	3A2	3	31.1	69.3	30	α/γ-Proteobacteria/ Actinobacteria	*dmp*OPQ**B**, *dmp*C*DEFGH*	
	5F10	3	42.1	64.0	36	β/γ-Proteobacteria	*dmp***B**C*E*, *phe***B**	two EDO genes (I.2.A and I.2.G subfamilies) and presence of four tRNA genes
	2H2	1	37.5	64.9	30	β/γ-Proteobacteria	*dmp***B**C*E*	presence of four tRNA genes
	9B1	1	36.8	64.7	30	β/γ-Proteobacteria	*dmp***B**C*E*	presence of four tRNA genes
I.2.B	4D5	3	40.0	60.1	42	α-Proteobacteria	*dmp*Q**B**C*EFGH*	presence of genes for branched-chain amino-acid ABC transporter system
	1A9	2	31.5	60.0	32	α-Proteobacteria	*dmp*Q**B**C*EFGH*	id.[c]
	6F5	2	36.1	60.4	37	α-Proteobacteria	*dmp*Q**B**C*EFGH*	id.[c]
	10D8	3	35.2	64.1	32	α/β/γ-Proteobacteria	*dmp***B**C*DEFGH*	id.[c]
I.2.C	3F10	2	40.7	67.7	44	β/γ-Proteobacteria	*tdn*D1C-D2C2E*FGHIJKL*	two EDO genes (I.2.C and I.2.A subfamilies)
	9C3	1	34.4	65.6	37	β/γ-Proteobacteria	*tdn*QTA1A2BRD1-C*EFGLHIJ*	complete *tdn* gene cluster and presence of complete *mer* gene cluster
	7D12	1	37.2	66.0	44	β-Proteobacteria	*Dmp*OPQ**B**	presence of *rpsI* and *rplM* homologs, and Tn4371-like gene island
	4E8	2	41.3	66.0	45	β-Proteobacteria	*Dmp*PQ**B**	id.[c]
	6H11	1	31.5	66.0	38	β-Proteobacteria	*Dmp*PQ**B**	id.[c]
	9E5	1	32.8	66.0	39	β-Proteobacteria	*Dmp*OPQ**B**	id.[c]
I.1.C	6B9	2	34.3	62.8	26	α/γ-Proteobacteria	EDO gene[b], *dmp***B**	two EDO genes (I.1.C and I.2.A subfamilies)
	4A3	1	41.6	63.0	31	α/γ-Proteobacteria	EDO gene[b]	
I.3.M	7E11	1	36.8	66.7	34	Actinobacteria	EDO gene[b]	
	8C3	2	34.7	66.6	32	Actinobacteria	EDO gene[b]	
I.3.N	1F2	1	39.8	64.1	36	α-Proteobacteria	*nah*NCLOMG, *nag*GI	
I.2.G	1D9	1	36.8	60.8	34	α-Proteobacteria	*phe***AB**	complete pSKYE1
	1D2	1	34.4	60.8	34	α-Proteobacteria	*phe***AB**	a part of pSKYE1
	9B9	1	36.7	60.7	35	α-Proteobacteria	*phe***AB**	complete pSKYE1
	3H5	1	36.7	60.7	35	α-Proteobacteria	*phe***AB**	complete pSKYE1
	1H11	1	34.5	60.8	32	α-Proteobacteria	*phe***AB**	a part of pSKYE1
	6D4	2	30.1	61.2	34	α-Proteobacteria	*phe***AB**	a part of pSKYE1
	1E12	1	37.4	60.7	35	α-Proteobacteria	*phe***AB**	
	5F2	1	34.7	62.8	33	α-Proteobacteria	*phe***AB**CC	
	9C8	1	35.4	62.8	29	α-Proteobacteria	*phe***AB**CC	Presence of putative genes for conjugation of ICE/plasmid
	1A1	1	35.0	62.8	31	α-Proteobacteria	*phe***AB**CC	id.[c]
	3G3	3	41.3	63.3	35	α-Proteobacteria	*phe***AB**	id.[c]
	9E4	2	37.8	61.9	36	α-Proteobacteria	*phe***AB**CC, *dmp***B**	Two EDO genes (I.2.G and I.2.A subfamilies)

Table 1 Continued

EDO subfamily	Fosmid name	Number of contigs	Size (kb)	G+C content (%)	Number of ORFs	Phylogenic affiliation by SOM analysis	Putative degradation genes[a]	Remarks
	7B2	1	33.9	61.7	33	α-Proteobacteria	phe**ABCC**	
	4E12	1	36.3	63.3	36	α/β-Proteobacteria	phe**ABCC**	
	2B9	2	29.1	61.9	30	α-Proteobacteria	phe**ABC**	
	2A1	1	33.5	61.5	34	α-Proteobacteria	phe**ABCC**	
	5B2	1	33.8	60.9	36	α-Proteobacteria	phe**ABCC**, phe**A2**	
	2C1	2	33.0	61.9	41	α-Proteobacteria	phe**ABCC**, phe**A2A1**, dmp**EFGH**	

[a]EDO genes are shown in bold.
[b]No homology with any representative EDO genes, dmpB, tdnC, nagC, nahH, or pheB.
[c]id., idem.

regulator of the biphenyl-catabolizing (*bph*) operon of *P. pseudoalcaligenes* KF707 (Watanabe *et al.*, 2000). Fosmid types 3A2 and 2C5 differ remarkably in that the two *dmp* gene clusters in 3A2 are separated by an intervening 10-kb sequence that does not appear to be involved in the degradation of aromatic compounds. Therefore, the 3A2-type gene cluster may be an evolutionarily intermediate form of the 2C5-type complete operon. On the other hand, some other *dmp* genes may be located upstream of ORF3A2-1.

Fosmids 5F10, 2H2 and 9B1 have some DNA fragments in common (Figure 1 and Supplementary Figure S1). The nucleotide sequence of ORF5F10-3-encoding catechol 2,3-dioxygenase is identical to those of ORF26 in 6B9 (Figure 1 and Supplementary Figure S2) and ORF36 in 9E4 (Figure 1 and Supplementary Figure S3). The deduced gene products for ORF5F10-3, −5 and −6 show 56%, 60% and 66% identity to DmpB, DmpC and DmpE, respectively, of pVI150. In these fosmids, only these three ORFs were predicted to be involved in aromatics degradation. 5F10 contains four tRNA genes (for Pro, Arg, His and Lys), indicating that this DNA fragment is most probably of chromosomal origin (Figure 1).

Fosmids carrying I.2.B EDO genes. EDOs belonging to subfamily I.2.B are frequently identified in the genus *Sphingomonas*. They include catechol 2,3-dioxygenases, involved in the degradation of single-ring aromatic compounds (for example, xylenes) (Eltis and Bolin, 1996). I.2.B EDO genes were borne on four fosmids, three of which (4D5, 1A9 and 6F5) had nearly an identical sequence; with the exception that 4D5 carried two I.2.B EDO genes (ORF4D5-30 and −31) (Figure 1 and Supplementary Figure S1).

The deduced amino acid sequences of the ORF4D5-10 to −17 gene products are 37−70% identical to those of proteins involved in the degradation of phenol (DmpQBCEFGH of pVI150) and aniline (TdnD2C2-Orf4 product-TdnEGIJK of pTDN1 from *P. putida* UCC2) (Figure 1). However, no genes for putative phenol hydroxylase homologs were identified in the flanking regions, suggesting that the degradation gene cluster in 4D5 constitutes only part of a complete degradative pathway and is likely to work in concert with other upstream enzyme genes, which might be located at a distant locus in the same genome or provided by other organisms.

Fosmid 10D8 carries seven ORFs, ORF10D8-21, −22, −23, −26, −27, −28, −30, which encode polypeptides calculated to be 37−70% identical to the gene products of *dmpBCEHFGD* of pVI150 and *tdnC2*-Orf4-*tdnEGKIJF* of pTDN1 (Figure 1). ORF10D8-24 was predicted to be a GntR-type transcriptional regulator. Unlike other well-studied *meta*-pathway gene clusters, this putative regulator gene is positioned in the middle of the degradation gene cluster.

Table 2 The predicted functions of genes located at the flanking regions of EDO genes in fosmid clones

Group of fosmid clones[a]	Number of genes with predicted functions (%)							
	Deg.	Reg.	Trans.	Rep.	Conj.	DNA tra.	Others	Unknown
I.2.A (5)	37 (24)	8 (5)	2 (1)	12 (8)	0 (0)	5 (3)	37 (24)	56 (35)
I.2.B (4)	27 (19)	16 (11)	39 (28)	0 (0)	0 (0)	1 (1)	27 (19)	31 (22)
I.2.C (6)	41 (17)	11 (4)	5 (2)	20 (8)	5 (2)	23 (9)	74 (30)	68 (28)
I.1.C (2)	3 (5)	3(5)	10 (18)	3 (5)	0 (0)	0 (0)	30 (53)	8 (14)
I.3.M (2)	2 (3)	2 (3)	6 (9)	2 (3)	0 (0)	0 (0)	36 (55)	18 (27)
I.3.N (1)	15 (41)	4 (11)	6 (17)	0 (0)	0 (0)	0 (0)	9 (25)	2 (6)
I.2.G (18)	70 (11)	48 (8)	18 (3)	44 (7)	45 (7)	128 (21)	99 (16)	161 (27)
Total (38)	195 (15)	91 (7)	86 (7)	81 (6)	50 (4)	157 (12)	312 (23)	344 (26)

Abbreviations: Conj., conjugative transfer; Deg., degradation for aromatic compounds; DNA tra., DNA transposition and site-specific recombination; others, other known functions; Reg., transcriptional regulation; Rep., replication, maintenance, and processing of DNA; Trans., transport.
[a]Same grouping with Table 1. Scores in parentheses are numbers of analyzed DNA fragments.
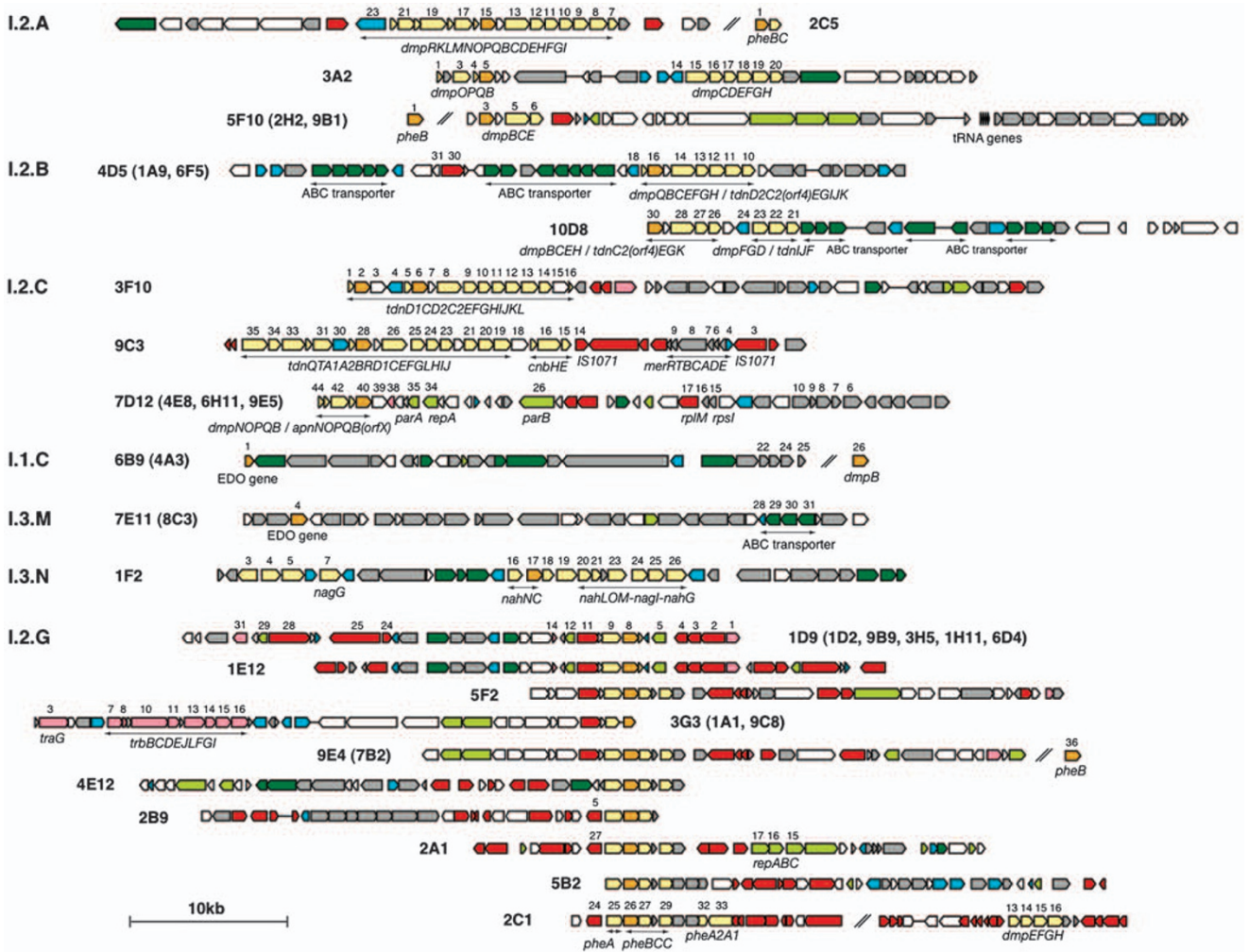

Figure 1 Genetic maps of representative fosmid inserts analyzed in this study. ORFs are indicated by pentagons, and those referred to in the text are shown with their numbers above the symbols. Predicted ORF functions are shown according to color: dark yellow, EDO; light yellow, degradation of aromatic compounds; green, replication, maintenance and processing of DNA; magenta, conjugative transfer; red, DNA transposition or site-specific recombination; dark green, transport; light blue, transcriptional regulation; gray, other known functions and white, unknown function. Predicted gene names for aromatic compound degradation and other relevant functions (Table 1) are shown under the symbols. Horizontal lines indicate non-scaled spaces between genes. Two contigs in one fosmid insert are separated by double oblique strokes. Fosmids carrying same inserts are described in parentheses.

Fosmids carrying I.2.C EDO genes. The EDOs of subfamily I.2.C are primarily catechol 2,3-dioxygenases responsible for the degradation of single-ring aromatic compounds (for example, phenol, nitrobenzene and aniline). This type of EDO is often found in β-*Proteobacteria* (*Comamonas*, *Ralstonia*
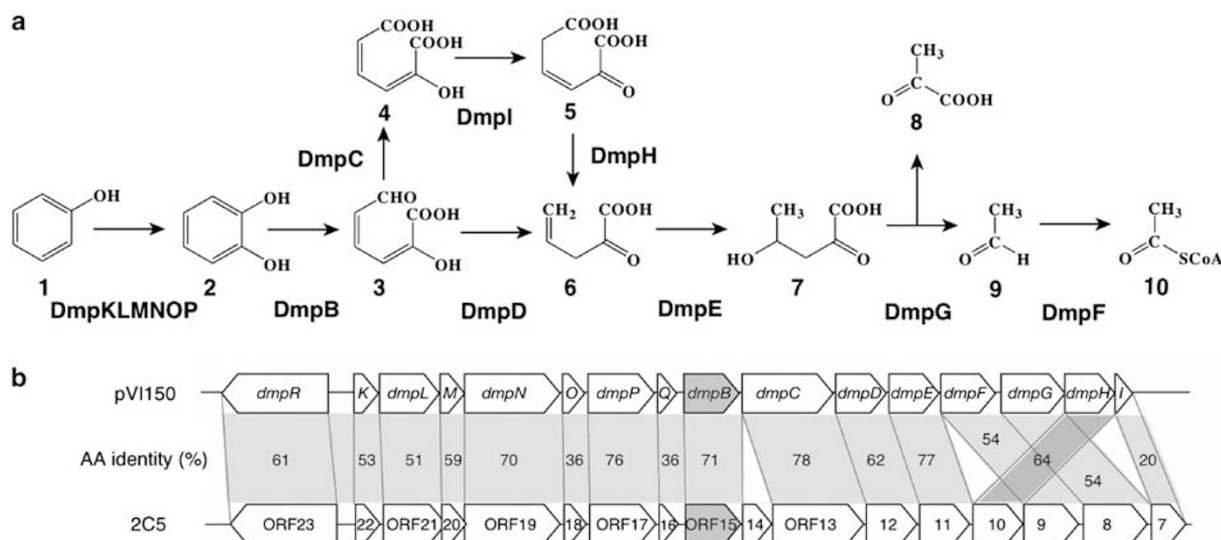

Figure 2 Genetic organization of phenol degradation gene clusters. (**a**) Phenol catabolism pathway of pVI150 from *Pseudomonas* sp. CF600. Proteins: DmpKLMNOP, components of phenol hydroxylase; DmpB, catechol 2,3-dioxygenase; DmpC, 2-hydroxymuconic semialdehyde dehydrogenase; DmpD, 2-hydroxymuconic semialdehyde hydrolase, DmpE, 2-oxopent-4-dienoate hydratase; DmpF, acetaldehyde dehydrogenase; DmpG, 4-hydroxy-2-oxovalerate aldolase; DmpH, 4-oxalocrotonate decarboxylase; DmpI, 4-oxalocrotonate isomerase; and DmpQ, unknown protein. Metabolic compounds: (1) phenol; (2) catechol; (3) 2-hydroxymuconic semialdehyde; (4) 4-oxalocrotonate (enol form of 2-hydroxyhexa-2,4-diene-1,6-dioate); (5) 4-oxalocrotonate (keto form of 2-oxohex-3-ene-1,6-dioate); (6) 2-oxopent-4-dienoate; (7) 4-hydroxy-2-oxovalerate; (8) acetaldehyde; (9) acetaldehyde; (10) acetyl-CoA. (**b**) Comparison of the pVI150-specified *dmp* cluster with the phenol-degrading gene cluster in fosmid 2C5. Open arrows indicate ORFs with their directions of transcription.
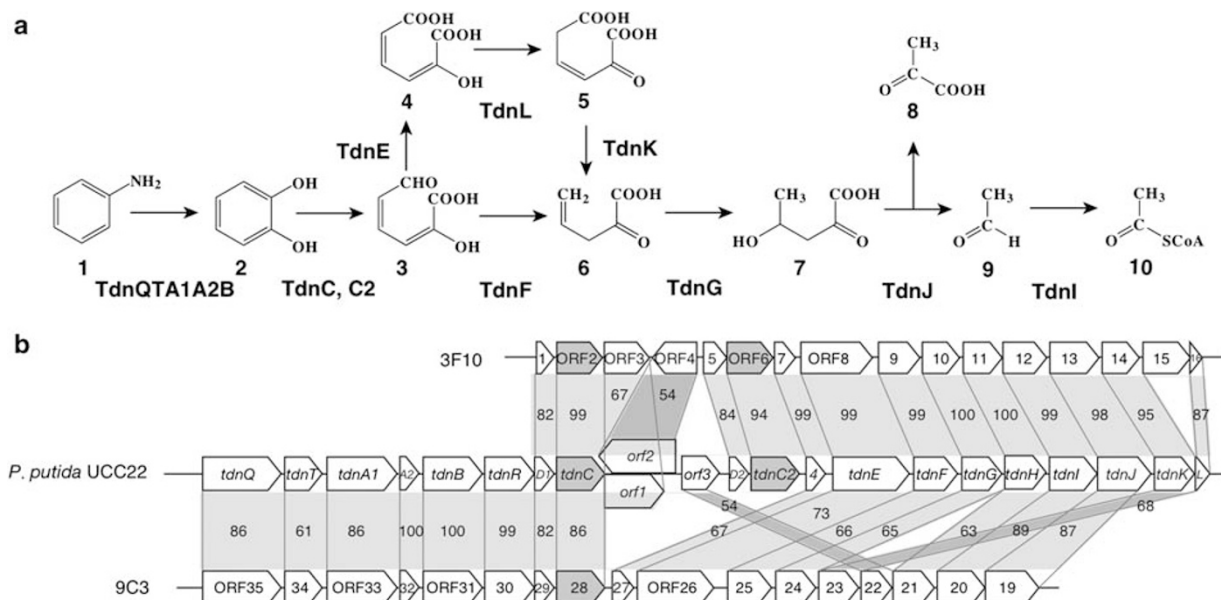

Figure 3 Genetic organization of aniline-degradation gene clusters. (**a**) Aniline catabolism pathway of pTDN1 from *P. putida* UCC22. Proteins: TdnQTA1A2B, components of aniline dioxygenase; TdnC and TdnC2, catechol 2,3-dioxygenases; TdnE, 2-hydroxymuconic semialdehyde dehydrogenase; TdnF, 2-hydroxymuconic semialdehyde hydrolase; TdnG, 2-oxopent-4-dienoate hydratase; TdnI, acetaldehyde dehydrogenase; TdnJ, 4-hydroxy-2-oxovalerate aldolase; TdnK, 4-oxalocrotonate decarboxylase; TdnL, 4-oxalocrotonate isomerase and TdnD1 and TdnD2; putative ferredoxins with undefined degradation function. Metabolic compounds: (1) aniline; (2) catechol; (3) 2-hydroxymuconic semialdehyde; (4) 4-oxalocrotonate, (enol form of 2-hydroxyhexa-2,4-diene-1,6-dioate); (5) 4-oxalocrotonate (keto form of 2-oxohex-3-ene-1,6-dioate); (6) 2-oxopent-4-dienoate; (7) 4-hydroxy-2-oxovalerate; (8) pyruvate; (9) acetaldehyde; (10) acetyl-CoA. (**b**) Comparison of pTDN1-specified *tdn* gene cluster with aniline-degradation gene clusters in 3F10 and 9C3. Open arrows show ORFs with their directions of transcription.
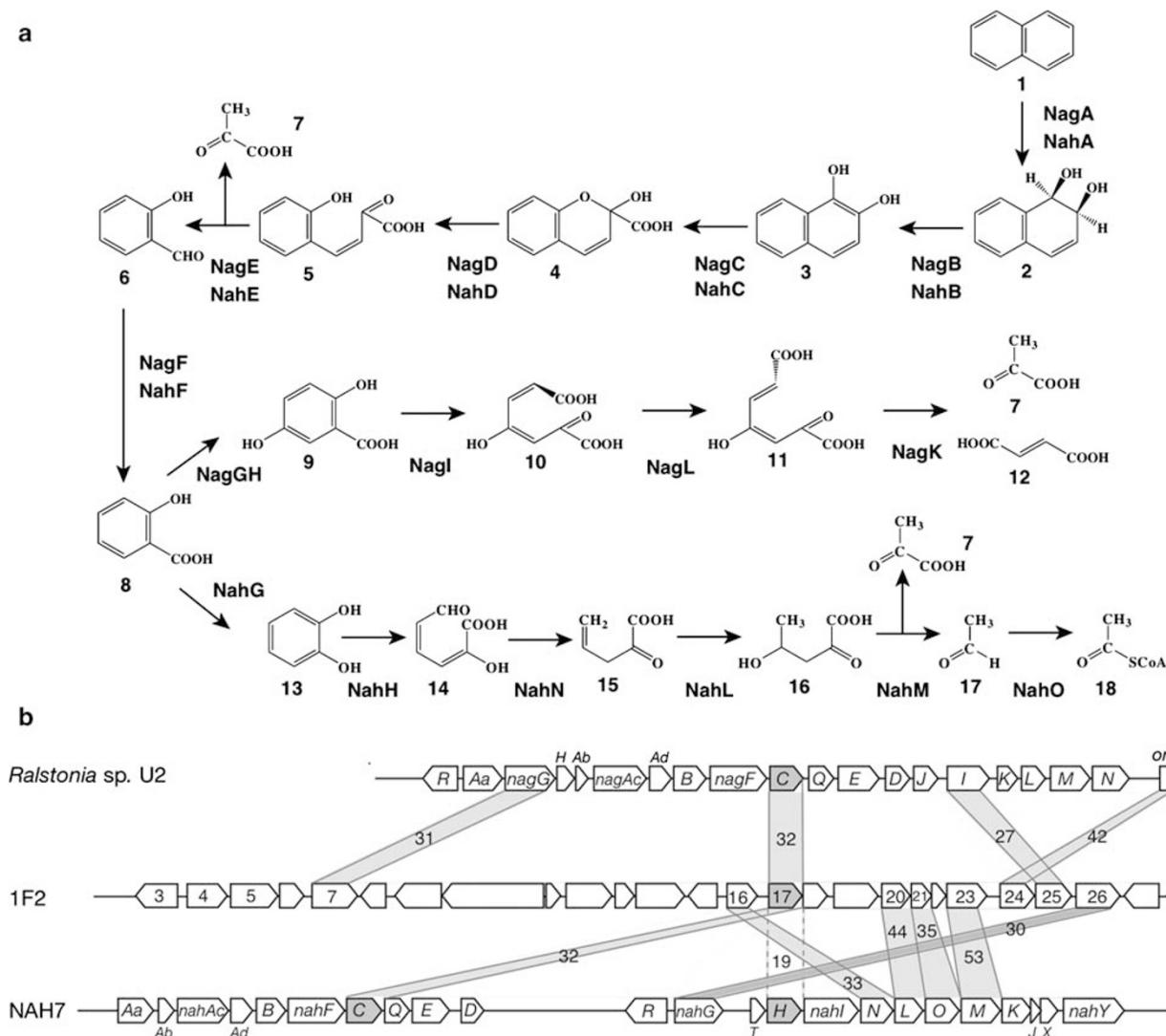
Figure 4 Genetic organization of naphthalene-degradation gene clusters. (**a**) Naphthalene catabolism pathways of Nah enzymes encoded by *NAH7* in *P. putida* G7 and Nag enzymes in *Ralstonia* sp. U2. Proteins: NagA and NahA, naphthalene dioxygenase; NagB and NahB, *cis*-1,2-dihydro-1,2-dihydroxynaphtalene-1,2-dehydrogenase; NagC and NahC, 1,2-dihydroxynaphthalene dioxygenase; NagD and NahD, 2-hydroxychromene-2-carboxylate isomerase; NagE and NahE, *trans-o*-hydroxybenzylidenepyruvate hydratase-aldolase; NagF and NahF, salicylaldehyde dehydrogenase; NagG, salicylate-5-hydroxylase component; NagH, salicylate-5-hydroxylase component; NagL, maleylpyruvate isomerase; NagK, fumarylpyruvate hydrolase, NahG, salicylate 1-hydroxylase; NahN, 2-hydroxymuconic semialdehyde hydrolase; NahL, 2-oxopent-4-dienoate hydratase; NahM, 4-hydroxy-2-oxovalerate aldolase; and NahO, acetaldehyde dehydrogenase. Metabolic compounds: (1) naphthalene; (2) *cis*-1,2-dihydro-1,2-dihydroxynaphtalene; (3) 1,2-dihydroxynaphthalene; (4) 2-hydroxychromene-2-carboxylate; (5) *cis-o*-hydroxybenzalpyruvate; (6) salicylaldehyde; (7) pyruvate; (8) salicylate; (9) gentitase; (10) maleylpyruvate; (11) fumarylpyruvate; (12) fumarate; (13) catechol; (14) 2-hydroxymuconic semialdehyde; (15) 2-oxopent-4-dienoate; (16) 4-hydroxy-2-oxovalerate; (17) acetaldehyde; (18) acetyl-CoA. (**b**) Comparison of *nag* and *nah* gene clusters with naphthalene-degradation gene clusters in 1F2. Open arrows show ORFs with their directions of transcription.

and *Burkholderia*) and γ-Proteobacteria (*Pseudomonas*) (Eltis and Bolin, 1996). I.2.C EDO genes were found in all three of our metagenomic fosmid clones (3F10, 9C3 and 7D12).

The pTDN1-specified genes in the aniline-degrading *meta*-cleavage pathway constitute the operon *tdnD1C(orf123)D2C2(orf4)EFGHIJKL* (Fukumori and Saint, 2001) (Figure 3). Most of the deduced proteins encoded by ORF3F10-1 to −16 show extremely high amino acid identity (>94%) with these Tdn proteins (Figures 1 and 3). As is the case for pTDN1, 3F10 has two genes each for catechol 2,3-dioxygen-

ase (ORF3F10-2 and −6) and ferredoxin (ORF3F10-1 and −5).

The deduced amino acid sequences of ORF9C3-19 to −35 are 54–100% identical to those of proteins encoded by pTDN1 (Figures 1 and 3). Unlike 3F10, the 9C3 fosmid contains single copies of the genes for catechol 2,3-dioxygenase and ferredoxin. To examine the activities of the ORF9C3-28 to −35 gene products, the gene cluster was subcloned into pUC18. *E. coli* JM109 cells carrying the recombinant plasmid produced 2-hydroxymuconic semialdehyde (detected by yellow coloration) on LB agar plates
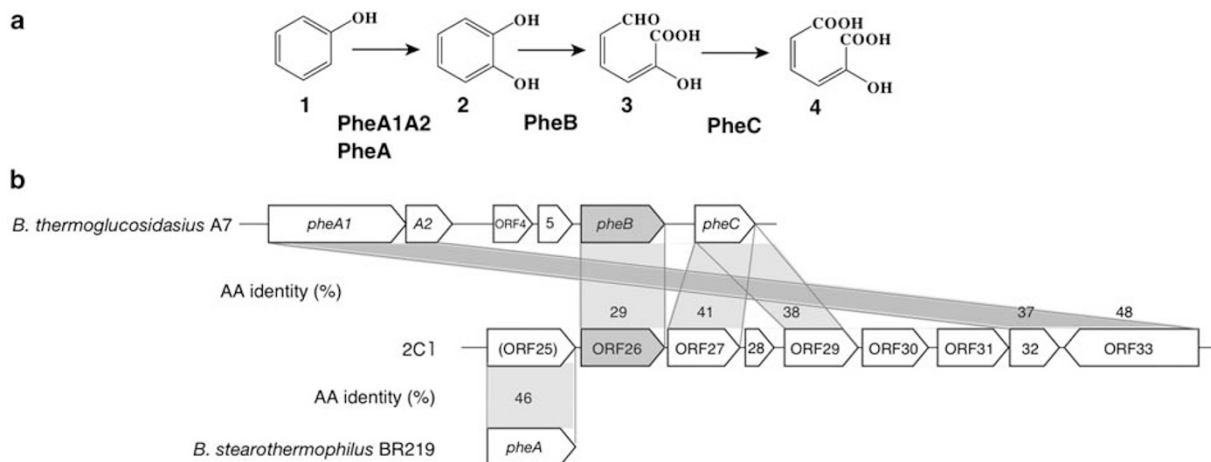
Figure 5 Part of the phenol degradation pathway and organization of its gene clusters in Gram-positive bacteria. (**a**) Part of the catabolic pathway for the degradation of phenol by Phe enzymes in two phenol-utilizing strains, *B. thermoglucosidasius* A7 and *B. stearothermophilus* BR219. Proteins: PheA (BR219) and PheA1A2 (A7), component(s) of phenol hydroxylase; PheB, catechol 2,3-dioxygenase; PheC, 2-hydroxymuconic semialdehyde dehydrogenase. Metabolic compounds: (1) phenol; (2) catechol; (3) 2-hydroxymuconic semialdehyde; (4) 4-oxalocrotonate (enol form of 2-hydroxyhexa-2,4-diene-1,6-dioate). (**b**) Comparison of phenol degradation gene cluster on the 2C1 insert with the *phe* gene clusters of A7 and BR219. Open arrows show the ORFs with their directions of transcription. Note that identity to the BR219 PheA product was determined using the amino acid sequence of the ORF1D2-9 product because of a frame-shift mutation in ORF2C1-25.

containing aniline, confirming the activities of aniline dioxygenase and catechol 2,3-dioxygenase.

The aniline-catabolizing gene cluster on pTDN1 is flanked by IS*1071* (Fukumori and Saint, 2001). IS*1071* is preferentially located on broad host-range IncP-1β plasmids and can transpose in β-proteobacterial strains with high frequency (Sota *et al.*, 2006b). Downstream of the degradation gene cluster on 9C3, two truncated IS*1071* transposase genes (ORF9C3-3 and −14) flank a complete set of mercury-resistance genes, *merRTPCADE* (ORF9C3-4 to 10) (Figure 1). Because IncP-1β plasmids often carry mercury-resistance genes flanked by IS*1071*, the insert in 9C3 might have originated from an IncP-1β plasmid of β-proteobacterial origin, which is consistent with the SOM analysis result.

Four fosmids (7D12, 4E8, 6H11 and 9E5) had overlapping sequences (Figure1 and Supplementary Figure S1). The deduced products of ORF7D12-39 to −44 show amino acid sequence identities of 71–83% with those of OrfX (unknown function), AphB (2,3-dihydroxybiphenyl dioxygenase), AphQ (unknown function), AphP (reductase component of phenol hydroxylase), AphO (hydroxylase component of phenol hydroxylase) and AphN (hydroxylase component of phenol hydroxylase) of the phenol-utilizing strain *Comamonas testosteroni* TA441 (Arai *et al.*, 2000); these levels of identity are higher than those found for the DmpOPQB products of pVI150, which exhibit only 33–58% identity with these proteins.

Fosmid 7D12 contained two genes for ribosomal proteins, L13 (RplM, ORF7D12-15) and S9 (RpsI, ORF7D12-16); L13 and S9 are 90% and 87% identical, respectively, to a ribosomal protein of *Acidvorax* sp. JS42, indicating that this DNA

fragment must have originated from a β-proteobacterial chromosome (Figure 1). ORF7D12-17, a putative gene (*int*) for tyrosine recombinase-type integrase that has highly conserved functional homologs in the Tn*4371*-like genomic islands (Toussaint *et al.*, 2003), was located downstream of *rplM*. Likewise, the deduced amino acid sequences of the ORF7D12-26, −34, −35 and −38 gene products show 90% identity with those of ParB, RepA, ParA and TraF, respectively, of the Tn*4371*-like islands. Furthermore, a conserved target site (5′-TTTTCAT-3′) for Tn*4371* insertion was found in the intergenic region between the *rplM* and *int* genes, which might constitute one end of the Tn*4371*-like genomic islands. Therefore, the 22-kb region downstream of *rplM* is likely to be part of a Tn*4371*-like genomic island that contains the *aph* operon.

Fosmids carrying I.1.C and I.3.M EDO genes. We recently proposed the I.1.C and I.3.M EDO subfamilies on the basis of their novel amino acid sequences (Suenaga *et al.*, 2007). A DNA fragment containing the I.1.C EDO gene was found in both fosmids 4A3 and 6B9, and a fragment containing the I.3.M EDO gene was found in both 7E11 and 8C3 (Figure 1 and Supplementary Figure S2). Strikingly, degradation genes other than these EDO genes were not identified in the fosmid inserts. ORF7E11-1, −2 and −3 appear to constitute an operon with the I.3.M EDO gene (ORF7E11-4). The former three genes were annotated as the genes for a hypothetical protein, hydantoin racemase and fumarate reductase flavoprotein subunit, respectively, but functions of these ORFs for the degradation of aromatic compounds remain unclear. Although there are some examples of functionally related catabolic

1344

genes being dispersed on chromosomes and plasmids (for example, *Rhodococcus* spp. and sphingomonads) (van der Geize and Dijkhuizen, 2004; Basta *et al.*, 2005), the 'solitary'-type gene structure identified in these fosmids has not previously been documented.

Fosmids carrying I.3.N EDO genes. The EDO (ORF1F2-17 product) encoded by fosmid 1F2 belongs to the novel subfamily I.3.N (Suenaga *et al.*, 2007). Sixteen genes (ORF1F2-3 to −7 and −16 to −26) in the fosmid were categorized as involved in aromatics degradation (Figures 1 and 4). Of these 16 genes, nine (ORF1F2-7, −16, −17, −20, −21, −23 to −26) encoded products predicted to have 27–53% amino acid sequence identity with those of the enzymes for naphthalene degradation encoded by the *nah* genes of plasmid NAH7 from *P. putida* G7 (Sota *et al.*, 2006a) and the *nag* genes of *Ralstonia* sp. U2 (Zhou *et al.*, 2001). However, these 16 genes are apparently not sufficient to constitute a complete pathway for naphthalene degradation (Figure 4). Although the EDO sequence of ORF1F2-17 is only 19% identical to that of NAH7-encoded NahH (catechol 2,3-dioxygenase), the ORF1F2-17 gene product actually cleaves catechol at the *meta*-position (Suenaga *et al.*, 2007). Thus, the ORF1F2-16 to −26 products seem to constitute a *nah-nag* chimeric pathway for degradation of salicylate (compound 8 in Figure 4). *E. coli* EPI300 cells carrying the 1F2 fosmid grown in LB medium containing chloramphenicol were able to transform phenol, but not naphthalene, to its *meta*-cleavage yellow product. Despite these experimental results, no 1F2 fosmid genes were found to be homologs of known genes for phenol hydroxylase or its components. Therefore, the fosmid appears to contain novel phenol hydroxylase genes; the candidate genes are ORF1F2-3 (4-hydroxybenzoate 3-monoxygenase), ORF1F2-4 (ferredoxin), ORF1F2-5 (aromatic compound dioxygenase), ORF1F2-18 (reductase) and ORF1F2-19 (hydroxylase).

Fosmids carrying I.2.G EDO genes. In the 38 fosmids, genes of the novel I.2.G EDO subfamily were overrepresented, constituting 18 clones (Figure 1 and Supplementary Figure S3). (The complete sequences of functional EDO genes in the 3G3 fosmid were not obtained because of failure during contig assembly.) This dominance of the I.2.G subfamily suggests that this type of novel EDO may have an important role in our target environment (that is, coke plant wastewater habitat). All of the 18 fosmids (except for 3G3) carried a set of three genes encoding single-component-type phenol hydroxylase, catechol 2,3-dioxygenase and 2-hydroxymuconic semialdehyde hydrolase, in that order (Table 3). The 2-hydroxymuconic semialdehyde hydrolase genes on these fosmids were apparently either intact or truncated at the 3′-end. The fosmids containing an apparently intact 2-hydroxymuconic

Table 3 Structural diversity of three sequential genes (phenol hydroxylase, EDO, and 2-hydroxymuconic semialdehyde hydrolase) in fosmid inserts containing I.2.G subfamily EDO genes

Clones	Phenol hydroxylase activity and mutation sites[a]	EDO[e]	Hydrolase
1D9	+	group 2	3'-deletion
1D2	+	group 2	3'-deletion
9B9	+[G1009A (Ala337Thr)]	group 2	3'-deletion
3H5	+	group 2	3'-deletion
1H11	+	group 2	3'-deletion
6D4	+	group 2	3'-deletion
1E12	+	group 2	3'-deletion
5F2	+	group 2	full
9C8	+	group 1	full
1A1	+	group 1	full
3G3	unassembled	unassembled	unidentified
9E4	+	group 1	full
7B2	+[G127A[b]]	group 3	full
4E12	−[T160-[c], C644 T]	group 1	full
2B9	−[G1138A (Ala380Thr)]	group 4	full
2A1	−[C114A[d], T410C, C644 T]	group 5	full
5B2	−[T410C (Leu137Pro), C644T(Ala215Val)]	group 5	full
2C1	−[T160-[c], T583G, C644 T, G652C]	group 6	full

[a]+, active; −, inactive.
[b]Synonymous mutation.
[c]Frameshift mutation.
[d]Nonsense mutation.
[e]EDO genes were classified into six groups by their nucleotide sequences (Suenaga *et al.*, 2007).

semialdehyde hydrolase gene tended to carry an additional homologous copy downstream. Such genes, for example, ORF2C1-27 and −29 in fosmid 2C1, were 59% identical to each other in nucleotide sequence and their respective products were 38% and 41% identical, respectively, in amino acid sequence to 2-hydroxymuconic semialdehyde hydrolase (encoded by *pheC*) from the phenol-utilizing *B. thermoglucosidasius* strain A7 (Duffner *et al.*, 2000) (Figure 5).

All of the phenol hydroxylase genes upstream of the I.2.G EDO genes are very similar to one another, and their deduced products are 46% identical to PheA from *B. stearothermophilus* BR219, the only known single-component-type phenol hydroxylase (Kim and Oriel, 1995) (Figures 1 and 5). Fosmid 2C1 is different from the other 17 fosmids in that it additionally contains two putative genes for phenol hydroxylases. These genes (ORF2C1-33 and −32) are arranged in a tail-to-tail configuration, and their deduced products are 48 and 37% identical to PheA and PheB, respectively, from *B. thermoglucosidasius* A7 (Figures 1 and 5).

To examine the functions of these putative phenol hydroxylase genes, each DNA region containing the phenol hydroxylase and EDO genes was subcloned into pUC18. JM109 cells, separately carrying each of the 17 resulting plasmids, produced a yellow product on LB agar in the presence of catechol (Table 3), confirming catechol 2,3-dioxygenase

activities. In addition, 12 of the 17 clones produced 2-hydroxymuconic semialdehyde (yellow) on LB agar plates containing phenol. These active clones contain the genes for functional single-component phenol hydroxylases. Frame-shift (2C1, 4E12), non-sense (2A1), or substitution (2B9, 5B2) mutations in the remaining, inactive genes might have caused the loss or alteration of enzymatic activity (Table 3).

Fosmids carrying the I.2.G EDO genes were characterized by a relative abundance in mobile genetic elements and their remnants. It is also noteworthy that the transposase genes for IS*Pme1* (in 1D9 to 7B2; see below section '*Reconstitution of a plasmid-like circular DNA from fosmid sequences carrying I.2.G EDO genes*') and an IS*Sop9*-related element (ORF2B9-5, ORF2A1-27 and ORF2C1-24) were located very close to the single-component phenol hydroxylase genes (Figure 1 and Supplementary Figure S3). The fosmid 3G3 contains at least 10 genes predicted to be involved in the type IV conjugative transfer system: the gene for the coupling protein (TraG) and for the mating-pair formation apparatus (Trb) (Figure 1 and Supplementary Figure S3). The organization of these genes (*trbBCDEJLFGI*) in 3G3 resembles that found in the Tn*4371* genomic island (Toussaint *et al.*, 2003) rather than that found in self-transmissible plasmids (Schröder and Lanka, 2005). These genes exhibit ~60% identity to those of typical Tn*4371*-type genomic islands (Toussaint *et al.*, 2003) and much higher (~90%) identities to those found on the second chromosome of the α-proteobacterial strain *Ochrobactrum anthropi* ATCC 49188. Therefore, the transfer-related gene cluster in 3G3 appears to be part of a variant Tn*4371*-type genomic island in which several α-proteobacterial genomes also reside.

The deduced gene products of ORF2A1-17, −16 and −15 share 96%, 89% and 44% identity, respectively, with the replication proteins RepA (ZP_01441940), RepB (ZP_01441941) and RepC (ZP_01038470) of *Roseovarius* sp. strains HTCC2601 (RepA and B) and 217 (RepC). This type of *repABC* operon controls stable replication and partitioning of large and low-copy-number plasmids and has been found only in α-*Proteobacteria* (Cevallos *et al.*, 2008). Therefore, the 2A1 insert appears to originate from an α-proteobacterial plasmid, consistent with the result of SOM analysis.

Reconstitution of a plasmid-like circular DNA from fosmid sequences carrying I.2.G EDO genes
The fosmid fragments in 1D9, 1D2, 9B9, 3H5, 1H11 and 6D4 had overlapping sequences that were assembled into a 36 684-bp circular DNA form (Figure 6). This circular DNA, designated pSKYE1, contained a total of 34 ORFs and was considered to be a plasmid because of the observations described below.

The circular DNA pSKYE1 carries several genes related to plasmid replication (ORF1D9-5), partition
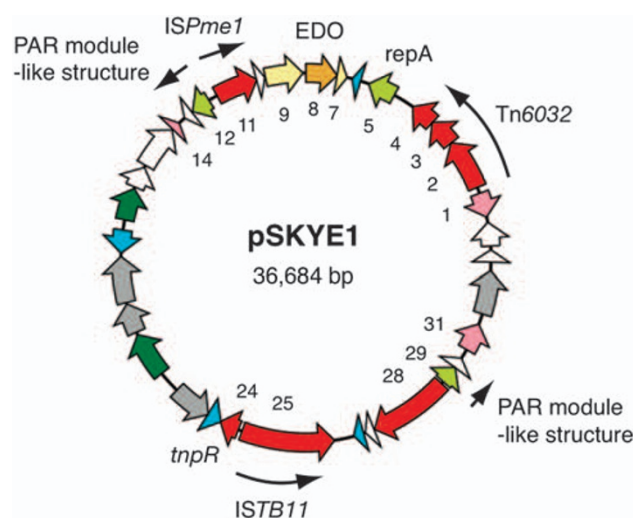

Figure 6 Circular map of plasmid pSKYE1 reconstructed *in silico*. The ORF numbers are those used for the fosmid 1D9 insert. Predicted ORF functions are shown according to color as in Figure 1.

(ORF1D9-12, and −29), and conjugation (ORF1D9-1, −14 and −31). The deduced gene product of ORF1D9-5 is 82% identical to RepA of plasmid pAMI2 of *Paracoccus aminophilus* JCM 7686, which has a replication system that is functional in several α-proteobacterial genera, such as *Paracoccus*, *Rhizobium*, *Rhodobacter* and *Agrobacterium* (Dziewit *et al.*, 2007). Therefore, replication of pSKYE1 might be possible in α-proteobacterial strains. The *repA* gene of pAMI2 is followed by a so-called PAR module, which is responsible for the active partitioning function of the plasmid (Dziewit *et al.*, 2007). Furthermore, pSKYE1 carries two putative partition gene modules (ORF1D9-12 and −13 and ORF1D2-29 and −30). The deduced gene products lack homology to the PAR modules in pAMI2. However, the deduced ORF1D9-12 and −13 gene products are 66% and 44% identical to the corresponding gene products encoded by *Laribacter hongkongensis* plasmid pHLHK22, and the deduced ORF1D2-29 and ORF1D9-30 gene products are 44% and 29% identical to the corresponding proteins encoded by a large plasmid from *Pseudomonas syringae* pv. *phaseolicola* 1448A. The deduced products of ORF1D9-1 and −14 are 70% and 87% identical to the C-terminal regions of the conjugative transfer TraG proteins of plasmid Ti from *Agrobacterium tumefaciens* (NP_059692) and plasmid pAgK84 from *A. tumefaciens* K84 (YP_086775), respectively, and the deduced ORF1D9-31 product is 51% identical to an intact conjugal transfer anti-restriction protein of *Nitrobacter hamburgensis* X14 (YP_575924).

An IS element and transposon were found in pSKYE1; their putative terminal inverted-repeat (IR) sequences are shown in Supplementary Figure S4. The former mobile genetic element flanked by 50-bp

1346

IR sequences is located very close to the single-component phenol hydroxylase gene and was thought to be IS*Pme1*, an IS element from plasmid pMTH1 of *Paracoccus methylutens* DM12 (Bartosik *et al.*, 2008). The latter element, designated Tn*6032*, has 25-bp IR sequences and carries three ORFs encoding polypeptides with moderate sequence identity to TniA (40%), TniB (35%) and an N-terminal portion of TniQ (23%) of Tn*5053* from *Xanthomonas* sp. W17 (Kholodii *et al.*, 1995). However, the IR sequences of Tn*6032* are very different from those of Tn*5053* and its related transposons, Tn*402*, Tn*5468* and Tn*7* (Oppon *et al.*, 1998; Minakhina *et al.*, 1999). Furthermore, the resolvase gene in Tn*5053* was not found in Tn*6032*. The 3.8-kb region containing ORF28 and ORF29 shows 90% nucleotide sequence identity to the corresponding region of the 3.8-kb Tn*3*-family transposon IS*TB11* on IncP-1α plasmid pTB11, which was exogenously isolated from a wastewater plant in Germany (Tennstedt *et al.*, 2005). The nucleotide sequence of IS*TB11* is distinct from that of other Tn*3*-family transposons, and its isoforms have not yet been discovered.

With respect to the genes involved in aromatic compound degradation, pSKYE1 contains only two genes for phenol hydroxylase and EDO. This situation is very different from that of known bacterial catabolic plasmids, which are usually large (>50 kb) and carry operonic degradative gene clusters (van der Meer *et al.*, 1992; Dennis, 2005). Because of their great lipophilicity and their ability to produce reactive oxygen species, phenol and catechol are potentially toxic to bacterial cells (Benndorf *et al.*, 2001). Wastewater from coke plants contains various organic pollutants, including phenol and mono- and polycyclic aromatic hydrocarbons (Stamoudis and Luthy, 1980). Therefore, pSKYE1, which carries a 'detoxification apparatus' for aromatic compounds, is likely to provide a survival advantage to its host cells in this harsh environment. The same environment might host bacterial strains that cannot degrade aromatic compounds by themselves. These bacteria may be fortunate in having this 'detoxification apparatus' functioning in their environment.

Conclusions

In our previous study (Suenaga *et al.*, 2007), we used a fosmid vector to clone environmental (that is, activated sludge) DNA fragments of 30–40 kb. These fragments were thought to be suitable for cloning large DNA fragments containing common aromatic compound degradation genes. In this study, we obtained various new insights into the distribution and organization of aromatics degradation genes in the activated sludge microbial community by analyzing the sequences of 38 selected clones. A major finding was that the complete pathways generally recognized as 'standard' in previous studies are extremely rare. Only two complete degradation pathways similar to *dmp* and *tdn* gene clusters were identified. Instead, various types of gene subsets were identified that were not similar to previously reported 'upper' or 'lower' pathway modules, suggesting that aromatic compounds in the sampled natural environment are degraded through the concerted actions of various fragmental pathways.

Strikingly, a small piece of the aromatics degradation pathway consisting of single-component type phenol hydroxylase and EDO was overrepresented. Taking into account the fact that this module is found on our plasmid-like circular DNA form pSKYE1, we conclude that pSKYE1 has a crucial role in the detoxification of aromatic compounds in our target environment. All of these findings were obtained from the analysis of nucleotides only 1.5 Mb in length. Thus, sequence analysis of metagenomes showing specific functions is extremely useful in uncovering novel findings concerning targeted biological functions and their evolutionary stages across a wide range of bacterial strains.

References

Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T. (2003). Informatics for unveiling hidden genome signatures. *Genome Res* **13**: 693–702.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Amann RI, Ludwig W, Schleifer KH. (1995). Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol Rev* **59**: 143–169.

Arai H, Ohishi T, Chang MY, Kudo T. (2000). Arrangement and regulation of the genes for *meta*-pathway enzymes required for degradation of phenol in *Comamonas testosteroni* TA441. *Microbiology* **146**: 1707–1715.

Bartosik D, Putyrski M, Dziewit L, Malewska E, Szymanik M, Jagiello E *et al.* (2008). Transposable modules generated by a single copy of insertion sequence IS*Pme1* and their influence on structure and evolution of natural plasmids of *Paracoccus methylutens* DM12. *J Bacteriol* **190**: 3306–3313.

Basta T, Buerger S, Stolz A. (2005). Structural and replicative diversity of large plasmids from sphingomonads that degrade polycyclic aromatic compounds and xenobiotics. *Microbiology* **151**: 2025–2037.

Benndorf D, Loffhagen N, Babel W. (2001). Protein synthesis patterns in *Acinetobacter calcoaceticus*

induced by phenol and catechol show specificities of responses to chemostress. *FEMS Microbiol Lett* **200**: 247–252.

Cevallos MA, Cervantes-Rivera R, Gutiérrez-Ríos RM. (2008). The *repABC* plasmid family. *Plasmid* **60**: 19–37.

Dennis JJ. (2005). The evolution of IncP catabolic plasmids. *Curr Opin Biotechnol* **16**: 291–298.

Duffner FM, Kirchner U, Bauer MP, Müller R. (2000). Phenol/cresol degradation by the thermophilic *Bacillus thermoglucosidasius* A7: Cloning and sequence analysis of five genes involved in the pathway. *Gene* **256**: 215–221.

Dziewit L, Jazurek M, Drewniak L, Baj J, Bartosik D. (2007). The SXT conjugative element and linear prophage N15 encode toxin-antitoxin-stabilizing systems homologous to the *tad-ata* module of the *Paracoccus aminophilus* plasmid pAMI2. *J Bacteriol* **189**: 1983–1997.

Eltis LD, Bolin JT. (1996). Evolutionary relationships among extradiol dioxygenases. *J Bacteriol* **178**: 5930–5937.

Esteve-Núñez A, Caballero A, Ramos JL. (2001). Biological degradation of 2,4,6-trinitrotoluene. *Microbiol Mol Biol Rev* **65**: 335–352.

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR *et al.* (2008). The Pfam protein families database. *Nucleic Acids Res* **36**: D281–D288.

Fukumori F, Saint CP. (2001). Complete nucleotide sequence of the catechol metabolic region of plasmid pTDN1. *J Gen Appl Microbiol* **47**: 329–333.

Furukawa K, Suenaga H, Goto M. (2004). Biphenyl dioxygenase: functional versatilities and directed evolution. *J Bacteriol* **186**: 5189–5196.

Galvao TC, Mohn WW, de Lorenzo V. (2005). Exploring the microbial biodegradation and biotransformation gene pool. *Trends Biotechnol* **23**: 497–506.

Gibson DT, Parales RE. (2000). Aromatic hydrocarbon dioxygenases in environmental biotechnology. *Curr Opin Biotechnol* **11**: 236–243.

Gibson J, Harwood CS. (2002). Metabolic diversity in aromatic compound utilization by anaerobic microbes. *Annu Rev Microbiol* **56**: 345–369.

Junca H, Pieper DH. (2004). Functional gene diversity analysis in BTEX contaminated soils by means of PCR-SSCP DNA fingerprinting: Comparative diversity assessment against bacterial isolates and PCR-DNA clone libraries. *Environ Microbiol* **6**: 95–110.

Kholodii GY, Mindlin SZ, Bass IA, Yurieva OV, Minakhina SV, Nikiforov VG. (1995). Four genes, two ends, and a *res* region are involved in transposition of Tn*5053*: a paradigm for a novel family of transposons carrying either a *mer* operon or an integron. *Mol Microbiol* **17**: 1189–1200.

Kim IC, Oriel PJ. (1995). Characterization of the *Bacillus stearothermophilus* BR219 phenol hydroxylase gene. *Appl Environ Microbiol* **61**: 1252–1256.

Lowe TM, Eddy SR. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**: 955–964.

Minakhina S, Kholodii G, Mindlin S, Yurieva O, Nikiforov V. (1999). Tn*5053* family transposons are *res* site hunters sensing plasmidal *res* sites occupied by cognate resolvases. *Mol Microbiol* **33**: 1059–1068.

Mohn WW, Garmendia J, Galvao TC, de Lorenzo V. (2006). Surveying biotransformations with *à la carte* genetic

traps: translating dehydrochlorination of lindane (gamma-hexachlorocyclohexane) into *lacZ*-based phenotypes. *Environ Micobiol* **8**: 546–555.

Nojiri H, Shintani M, Omori T. (2004). Divergence of mobile genetic elements involved in the distribution of xenobiotic-catabolic capacity. *Appl Microbiol Biotechnol* **64**: 154–174.

Ohtsubo Y, Ikeda-Ohtsubo W, Nagata Y, Tsuda M. (2008). GenomeMatcher: a graphical user interface for DNA sequence comparison. *BMC Bioinformatics* **9**: 376.

Oppon JC, Sarnovsky RJ, Craig NL, Rawlings DE. (1998). A Tn*7*-like transposon is present in the *glmUS* region of the obligately chemoautolithotrophic bacterium *Thiobacillus ferrooxidans*. *J Bacteriol* **180**: 3007–3012.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R *et al.* (2005). InterProScan: protein domains identifier. *Nucleic Acids Res* **33**: W116–W120.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA *et al.* (2000). Artemis: sequence visualization and annotation. *Bioinformatics* **16**: 944–945.

Sambrook J, Russell DW. (2001). *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Press, Cold Spring Harbor: New York, USA.

Schröder G, Lanka E. (2005). The mating pair formation system of conjugative plasmids—A versatile secretion machinery for transfer of proteins and DNA. *Plasmid* **54**: 1–25.

Shingler V, Powlowski J, Marklund U. (1992). Nucleotide sequence and functional analysis of the complete phenol/3,4-dimethyl phenol catabolic pathway of *Pseudomonas* sp. strain CF600. *J Bacteriol* **174**: 711–724.

Sipilä TP, Keskinen A-K, Åkerman M-L, Fortelius C, Haahtela K, Yrjälä K. (2008). High aromatic ring-cleavage diversity in birch rhizosphere: PAH treatment-specific changes of I.E.3 group extradiol dioxygenases and 16S rRNA bacterial communities in soil. *ISME J* **2**: 968–981.

Sota M, Yano H, Ono A, Miyazaki R, Ishii H, Genka H *et al.* (2006a). Genomic and functional analysis of the IncP-9 naphthalene-catabolic plasmid NAH7 and its transposon Tn*4655* suggests catabolic gene spread by a tyrosine recombinase. *J Bacteriol* **188**: 4057–4067.

Sota M, Yano H, Nagata Y, Ohtsubo Y, Genka H, Anbutsu H *et al.* (2006b). Functional analysis of unique class II insertion sequence IS*1071*. *Appl Environ Microbiol* **72**: 291–297.

Stamoudis VC, Luthy RG. (1980). Determination of biological removal of organic constituents in quench waters from high-BTU coal-gasification pilot plants. *Water Res* **14**: 1143–1156.

Suenaga H, Ohnuki T, Miyazaki K. (2007). Functional screening of a metagenomic library for genes involved in microbial degradation of aromatic compounds. *Environ Microbiol* **9**: 2289–2297.

Tennstedt T, Szczepanowski R, Krahn I, Pühler A, Schlüter A. (2005). Sequence of the 68 869 bp IncP-1α plasmid pTB11 from a waste-water treatment plant reveals a highly conserved backbone, a Tn*402*-like integron and other transposable elements. *Plasmid* **53**: 218–238.

Top EM, Springael D. (2003). The role of mobile genetic elements in bacterial adaptation to xenobiotic organic compounds. *Curr Opin Biotechnol* **14**: 262–269.

Toussaint A, Merlin C, Monchy S, Benotmane MA, Leplae R, Mergeay M *et al.* (2003). The biphenyl-

1348

and 4-chlorobiphenyl-catabolic transposon Tn*4371*, a member of a new family of genomic islands related to IncP and Ti plasmids. *Appl Environ Microbiol* **69**: 4837–4845.

Tsuda M, Tan HM, Nishi A, Furukawa K. (1999). Mobile catabolic genes in bacteria. *J Biosci Bioeng* **87**: 401–410.

van der Geize R, Dijkhuizen L. (2004). Harnessing the catabolic diversity of rhodococci for environmental and biotechnological applications. *Curr Opin Microbiol* **7**: 255–261.

van der Meer JR, de Vos WM, Harayama S, Zehnder AJ. (1992). Molecular mechanisms of genetic adaptation to xenobiotic compounds. *Microbiol Rev* **56**: 677–694.

Watanabe T, Inoue R, Kimura N, Furukawa K. (2000). Versatile transcription of biphenyl catabolic *bph* operon in *Pseudomonas pseudoalcaligenes* KF707. *J Biol Chem* **275**: 31016–31023.

Wittich R-M. (1998). Degradation of dioxin-like compounds by microorganisms. *Appl Microbiol Biotechnol* **49**: 489–499.

Williams PA, Sayers JR. (1994). The evolution of pathways for aromatic hydrocarbon oxidation in *Pseudomonas*. *Biodegradation* **5**: 195–217.

Zhou N-Y, Fuenmayor SL, Williams PA. (2001). *nag* genes of *Ralstonia* (formerly *Pseudomonas*) sp. strain U2 encoding enzymes for gentisate catabolism. *J Bacteriol* **183**: 700–708.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)